An investigation of discrete-state discriminant approaches to single-sensor source separation

Valentin Emiya, Emmanuel Vincent, Rémi Gribonval

AN INVESTIGATION OF DISCRETE-STATE DISCRIMINANT APPROACHES TO SINGLE-SENSOR SOURCE SEPARATION

Valentin Emiya, Emmanuel Vincent, Rémi Gribonval

Équipe-projet METISS, INRIA-IRISA
Campus de Beaulieu, 35042 Rennes Cedex, France
{valentin.emiya, emmnanuel.vincent,remi.gribonval}@irisa.fr

ABSTRACT

This paper investigated a new scheme for single-sensor audio source separation. This framework is introduced comparatively to the existing Gaussian mixture model generative approach and is focusing on the mixture states rather than on the source states, resulting in a discrete, joint state discriminant approach. The study establishes the theoretical performance bounds of the proposed scheme and an actual source separation system is designed. The performance is computed on a set of musical recordings and a discussion is proposed, including the question of the source correlation and the possible drawbacks of the method.

Index Terms— Audio, source separation, single sensor

1. INTRODUCTION

Blind source separation (BSS) consists in analyzing a mixture and in estimating an approximate version of its components, or sources. The performance may then be measured by the so-called signal-to-distortion ratio (SDR) by comparing these estimates to the original source signals. While a number of approaches for BSS rely on independent component analysis [1], these techniques are either not satisfying or impossible to use in the case of single-sensor audio source separation. This problem – motivated by applications like automatic separation for karaoke, voice enhancement/noise reduction in hearing aids or in telephony – is addressed using spectral models based on discrete states, like Gaussian mixture models (GMM) [2] and hidden Markov models (MMC) [3], or on continuous states [4, 5]. In this paper, the discrete-state approach is investigated since it proved to be efficient [6], in particular in the Gaussian case [7].

The state-of-the-art discrete-state approach [2] is based on a generative source model. By focusing on modeling the separation between the components rather than their generation or their *a priori* distribution, discriminant approaches may outperform generative ones [8]. The current paper aims at studying a new discriminant discrete-state scheme for audio source separation. It consists in considering the possible states of the mixture, rather than the states of the sources, and the related separation filters.

Firstly, the theoretical performance bounds of the targeted approach are established, in a similar way as in previous studies on masking techniques [9]. It consists in finding the model parameters that maximize the SDR by assuming that the source signals are known. In the current case, we propose such an oracle estimator in order to obtain a diagnostics on the effects of the constraints introduced by the model.

Then, an actual source separation system based on the proposed scheme is introduced. It includes a training stage where the state-related parameters and the separation filters are estimated, and a decoding algorithm.

The paper is structured as follows: the principle of the state-of-the-art GMM system is summarized in Section 2; the proposed discriminant scheme is described in Section 3, including an introduction to its main principles (3.1), the design of the oracle estimator associated to the theoretical performance bounds (3.2) and the actual source separation system (3.3). Quantitative and comparative performance are drawn in Section 4 together with a discussion on the results and on the advantages and drawbacks of the approach. Conclusions are finally drawn in Section 5.

Besides, in the rest of the paper, we consider a single-sensor additive mixture $x_t = \sum_{j=1}^{J} s_{jt}$ of J sources s_{jt} with a sampling frequency f_s. In the time frequency domain, the short time Fourier transforms (STFT) of the mixture and of the sources are defined on a discrete set F of frequencies f and on a discrete set T of time stamps t; they are denoted by X_{tf} and S_{jtf} respectively. For clarity, the variation bounds for indexes j, t and f will be omitted when the full definition range is referred to.

2. DISCRETE-STATE SOURCE SEPARATION

2.1. Time-frequency masking

Time-frequency (TF) masking is a separation technique widely used in audio source separation and consists in estimating the STFT of source j as

$$\widehat{S_{jtf}} \triangleq \alpha_{jtf} X_{tf} \tag{1}$$

α_{jtf} being the TF mask to be estimated. Several cases may be considered, including: the general case $\alpha_{jtf} \in \mathbb{R}$; the case $\alpha_{jtf} \geq 0$ with $\sum_j \alpha_{jtf} = 1$ which is associated to Wiener filtering methods; and the binary mask case $\alpha_{jtf} \in \{0, 1\}$ which is appropriate when sources are not overlapping in the TF domain. The second case – the positive masks – is considered throughout this paper.

2.2. Factorial source separation with GMM

In GMM-based approaches [2], each source is allocated Q states[1]. State $q \in [\![1; Q]\!]$ of source j has an *a priori* probability π_{jq} and the power spectral density (PSD) of source j in state q at frequency f

[1] We here consider the same number of states for each source. The extension to source-dependent numbers of states is straightforward.

is denoted by σ^2_{jqf}. We obtain a Q-state GMM in which the likelihood related to state q is

$$S_{jtf}|q \sim \mathcal{N}\left(0, \sigma^2_{jqf}\right) \quad (2)$$

The learning stage thus consists in estimating the state parameters from isolated sources using the following algorithm:

Algorithm 1 GMM-based source separation: learning stage for source j (Expectation Maximization algorithm)

Require: learning set $\left\{X, \{S_j\}_j\right\}$
  loop
    Update posteriors: $\gamma_{jqt} \propto p\left(\{S_{jtf}\}_f|q\right)\ \pi_{jq}$ via eq.(2).
    Update priors: $\pi_{jq} \propto \sum_t \gamma_{jqt}$.
    Update source variances: $\sigma^2_{jqf} \leftarrow \frac{\sum_t \gamma_{qt}\ |S_{jtf}|^2}{\sum_t \gamma_{qt}}$.
  end loop
  return $\{\pi_{jq}\}$ and $\left\{\sigma^2_{jqf}\right\}$

In the separation stage, the observed mixture at time t results from one of the $K \triangleq Q^J$ underlying factorial states. In such a given joint state $k = (k_1, \ldots, k_J) \in [\![1; Q]\!]^J$, source j is in state k_j and the likelihood of the mixture is

$$X_{tf}|k \sim \mathcal{N}\left(0, \sum_j \sigma^2_{jk_jf}\right) \quad (3)$$

The estimate $\widehat{S_{jtf}}$ of the STFT of source j is obtained by TF masking, the mask being defined by

$$\alpha^{\text{GMM}}_{jtf} \triangleq \sum_k \gamma_{kt} \frac{\sigma^2_{jk_jf}}{\sum_{j'} \sigma^2_{j'k_{j'}f}} \quad (4)$$

where $\gamma_{kt} \triangleq p\left(k|\{X_{tf}\}_f\right)$ is the *a posteriori* probability of state k, which is computed using $\gamma_{kt} \propto p\left(\{X_{tf}\}_f|k\right)\prod_j \pi_{jk_j}$. The decoding algorithm is thus:

Algorithm 2 GMM-based source separation: decoding stage

Require: test signal X, learned parameters $\left\{\sigma^2_{jqf}\right\}$ and $\{\pi_{jq}\}$.
  Compute posteriors: $\gamma_{kt} \propto p\left(\{X_{tf}\}_f|k\right)\prod_j \pi_{jk_j}$ via eq.(3).
  Compute TF masks using eq. (4).
  return source estimates $\left\{\widehat{S_j}\right\}$.

3. DISCRETE-STATE DISCRIMINANT APPROACHES

3.1. Main ideas

As seen in the previous section, the GMM approach uses TF masks with a particular structure given by eq. (4). Let us consider a TF mask with a more general structure:

$$\alpha^{\text{DISC}}_{jtf} \triangleq \sum_k g_{kt} w_{jkf} \quad (5)$$

where $k \in [\![1; K]\!]$ is the index of one of the K *joint states*, g_{kt} is the activation coefficient of state k at time t such that $\sum_k g_{kt} = 1$ and $w_{jkf} \in [0; 1]$ is the separation filter related to state k and source j at frequency f such that $\sum_j w_{jkf} = 1$. Note that the TF mask given by eq. (4) in the GMM-based approach is a particular case of eq. (5) where $g_{kt} = \gamma_{kt}$ and where w_{jkf} is a function of the variances of the sources. In the general case, joint states k are not considered as factorial states and each w_{jkf} is a free parameter that is not associated to the variances (no more defined) of the sources.

The first main idea is thus to focus on the states of the mixture directly, rather than on the states of the sources. For a given performance level, one possible benefit is to decrease the overall complexity compared to a factorial approach: the latter models the sources before combining them in an exhaustive, high-computational way in the decoding stage. However, all the factorial states may not be useful, in particular when dealing with correlated sources. By selecting the K states that optimally model the mixture, the proposed approach may make it possible to save these useless joint states.

The second idea addresses the design of the separating filters w_{jkf}. Indeed, the generalized joint states are not related to a generative source model like in the GMM-based approach. The separating filters may then be computed in a way that they perform better separation than with the GMM approach. In particular, they may be obtained using a discriminant learning as explained below.

3.2. Theoretical performance bounds

As a generalization of the GMM case, the proposed approach theoretically outperforms the GMM approach. The theoretical bounds are established here, in order to quantify the performance gain due to the particular structure of the TF masks given by eq. (5).

The optimal parameters $\{\mathring{g}_{kt}, \mathring{w}_{jkf}\}$ are obtained by maximizing the SDR or quasi-equivalently by minimizing the square error:

$$\{\mathring{g}_{kt}, \mathring{w}_{jkf}\} = \underset{g_{kt}, w_{jkf}}{\arg\min} \sum_{jtkf} \left| S_{jtf} - \sum_k g_{kt} w_{jkf} X_{tf} \right|^2 \quad (6)$$

which is rewritten as a weighted non-negative matrix factorization (NMF)[2] problem:

$$\{\mathring{g}_{kt}, \mathring{w}_{jkf}\} = \underset{w_{jkf}}{\arg\min} \sum_{jtf} \left(\frac{\text{Re}\left(S_{jtf} X^*_{tf}\right)}{|X_{tf}|} - [GW_j]_{tf}\ |X_{tf}| \right)^2 \quad (7)$$

where $[W_j]_{kf} \triangleq w_{jkf}$ and $[G]_{tk} \triangleq g_{kt}$. Using the gradient of eq. (7), the parameters are obtained via the multiplicative update:

$$g_{kt} \leftarrow g_{kt} \frac{\sum_{jf} \text{Re}\left(S_{jtf} X^*_{tf}\right) w_{jkf}}{\sum_{jf} |X_{tf}|^2 w_{jkf} \sum_{k'} g_{k't} w_{jk'f}} \quad (8)$$

$$w_{jkf} \leftarrow w_{jkf} \frac{\sum_t \text{Re}\left(S_{jtf} X^*_{tf}\right) g_{kt}}{\sum_t |X_{tf}|^2 g_{kt} \sum_{k'} w_{jk'f} g_{k't}} \quad (9)$$

Note that in eq. (8) and (9), the numerators may be negative. In order to ensure a non-negative factorization, a minimum, non-negative threshold is used on the numerators.

The theoretical performance bounds are thus obtained by iteratively performing the updates above. Parameters g_{kt} and w_{jkf}

[2]Note that a close weighted NMF approach was developed in [10] in another context. However, it differs from the current approach on several aspects including the objective function, the models in use and the weights.

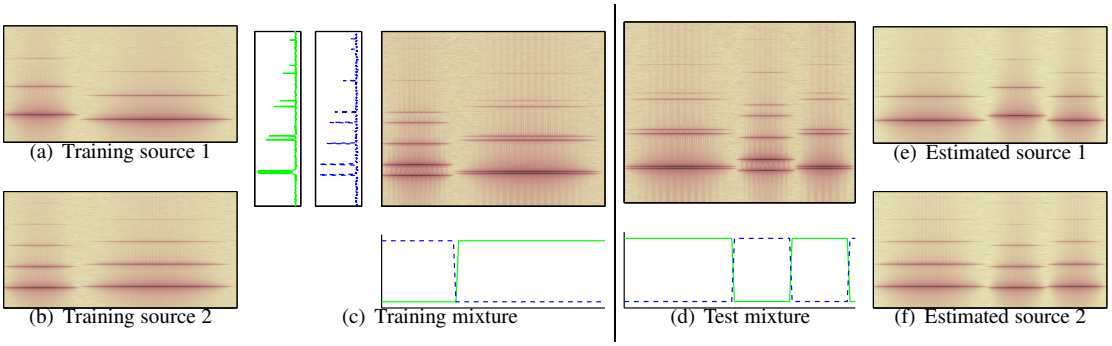

Figure 1: in a simple case, 2 sources (spectrograms (a) and (b)) are mixed to train a 2-joint-state model ((c), with learned DSPs vertically on the left and related posteriors horizontally down below). A test signal (d) is separated (estimated posterior down below) into (e) and (f).

may be initialized by random values, or by a generative GMM training on isolated sources. In the latter case, the performance improvement of the proposed approach is straightforward, since the multiplicative updates ensure that the SDR is not decreasing. This theoretical SDR gain will be quantified in Section 4.

3.3. Practical source separation system

The proposed joint-state framework may lead to several practical source separation systems that differ in their learning or decoding algorithms. We here introduce a first possible approach, which is illustrated in Fig. 1.

From one hand, the learning stage consists in finding K joint states and the associated prior π_k and PSD v_{kf} such that at time t,

$$X_{tf}|k \sim \mathcal{N}\left(0, v_{kf}\right). \tag{10}$$

We use an EM learning similar to the GMM-approach learning but applied to the mixtures instead of the sources, in order to select the most representative K states of the mixture. From the other hand, the learning of the separation filters is performed using the update given by eq. (9) at each the EM iteration[3]. The learning stage thus consists in the following algorithm applied to a learning database:

Algorithm 3 Proposed approach: learning stage

Require: learning set $\left\{X, \{S_j\}_j\right\}$.
  loop
    Update posteriors: $g_{kt} \propto p\left(\{X_{tf}\}_f|k\right)\ \pi_k$ using eq. (10).
    Update priors: $\pi_k \propto \sum_t \gamma_{kt}$.
    Update mixture variances: $v_{kf}^2 \leftarrow \frac{\sum_t \gamma_{kt}\left|X_{tf}\right|^2}{\sum_t \gamma_{kt}}$.
    Update separation filters W_{jkf} using eq. (9).
  end loop
  return $\{\pi_k\}$, $\{v_{kf}\}$ and $\{W_{jkf}\}$.

Finally, in a way similar to the GMM case, the decoding stage consists in the following algorithm:

[3]Note that we investigated another possible learning strategy based on the optimization given by eq. (6) and on both updates (8) and (9), but results have not been satisfying so far.

Algorithm 4 Proposed approach: decoding stage

Require: test signal X, learned parameters $\{\pi_k, v_{kf}, W_{jkf}\}$.
  Compute posteriors $g_{kt} \propto p\left(\{X_{tf}\}_f|k\right)\ \pi_k$ as a function of the observed STFT X and of the learned parameters v_{kf} and π_k.
  Estimate the STFT of the sources by TF masking using the learned separation filters $\widehat{w_{jkf}}$ in eq. (5).
  return source estimates $\left\{\widehat{S_j}\right\}$.

4. PERFORMANCE AND DISCUSSION

Theoretical bounds and blind separation performance is evaluated in the single-sensor case using a set of 10 musical recordings (available on request). The mixtures are composed by $J = 2$ sources, each piece being a song in which the singer and the accompaniment must be separated. All recordings are about 15 seconds long, sampled at 16kHz and the STFT is computed on 128ms-frames with 50% overlap.

In order to perform a comparative evaluation, the GMM system [2] is used as a state-of-the-art reference. The number of states per source is set to $Q \in \{1, 2, 4, 9, 16\}$. In the case of the proposed approach, we consider the number of joint states K since the isolated sources are not modeled. For comparison purposes, K is considered as equivalent to the number of factorial states of the GMM case, *i.e.* $K = Q^J \in \{1, 4, 16, 81, 256\}$. This equivalence – which may be discussed – is based on the complexity of the decoding stage rather than on the number of parameters to learn.

Theoretical performance is first drawn by assuming that separate sources are known. In the GMM case, models are separately trained on the source signals. The TF masks are then computed thanks to the *a posteriori* densities and to the variances of each source by combining them thanks to eq. (4). In the case of the proposed approach, the theoretical performance bounds detailed in Section 3.2 are computed using 100 iterations of eq. (8) and (9). In addition, we compute the TF masking oracle performance established in [9] by minimizing the quadratic error with respect to all the TF masks α_{jtf}'s without any discrete-state constraint – which is unrealistic for a BSS system. It should be seen as an upper bound that reflects the maximum separation performance due to the use of positive TF masks.

Blind separation performance is then evaluated. For each recording, the learning is performed on a 60s excerpt of the same musical piece at a different time location. This not very realistic

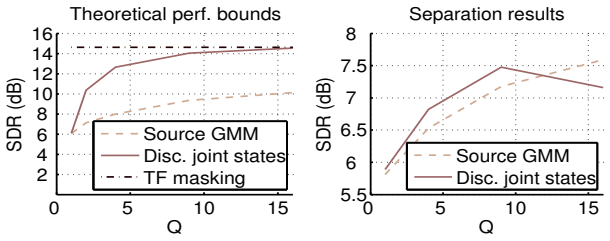

Figure 2: Theoretical (left) and separation (right) performance: the SDR is represented as a function of the equivalent number of states per source (with $K = Q^J$), for the GMM approach, the generalized, joint-state, discriminant approach and the optimal separation with non-constrained TF mask (from bottom to top).

but common evaluation procedure results in learning parameters that are well adapted to the mixture to be separated and makes it possible to compare several systems. Note that about 40 EM iterations are performed in the learning algorithms.

Results[4] are presented in Figure 2. On the left part, theoretical bounds confirm that the joint-state discriminant approach may outperform the generative, GMM-based approach. A significant SDR improvement of about 4 dB is obtained for $Q > 2$ and the theoretical bounds with the proposed approach are close to the TF mask upper bounds. Besides, the GMM performance SDR bound equals 10 dB for $Q = 16$ states per source while this SDR level is reached for $K = 4$ joint states (*i.e.* an equivalent $Q = 2$ value) in the proposed framework. This shows that in theory, the same performance may be reached by reducing the number of states and thus the computational cost.

This raises the question of the correlation between the sources. Indeed, sources are often assumed as decorrelated in BSS system for audio. This assumption may be valid in the case of speech signals but not for music signals in which source events often occurs in a synchronous way and harmonic structures are often related to a underlying tonality. We here observe that taking into account the source correlation may result in improving the separation results.

The right part of Figure 2 shows the results for actual source separation systems. Except for $Q = 16$ (*i.e.* $K = 256$), a small SDR improvement of about 0.3dB ($Q = 4$ and $Q = 9$) is obtained. This confirms that a source separation system can be designed in the proposed discriminant framework. Indeed, since the proposed modelization scheme is less constraining than in the GMM case, designing a related, actual separation system is not obvious once theoretical performance bounds are established (for instance, oracle performance associated to non-constrained TF masking are not realistic since no source separation system can be designed without adding any constraint).

However, the proposed separation system shows two limits. First, the system is sensitive to overfitting, which causes the poor SDR obtained for $Q = 16$ in Figure 2. This was additionaly checked thanks to an alternate training experiment: sources are randomly mixed resulting in a larger and more decorrelated training set; decoding results are then similar to the generative GMM case. Overfitting is a well-known drawback of discriminant approaches, which require larger training sets than generative approaches. This major concern should be specifically addressed in future works. As a second limit, the SDR improvements for $Q < 16$ is small, compared to the 4dB improvement obtained in the case of the theoretical performance bounds. Consequently, the proposed approach should be considered as a preliminary proof of concept rather than as an efficient source separation system, and other systems based on the proposed joint-state discriminant scheme may be investigated.

[4] Audio examples are available on the authors' website.

5. CONCLUSIONS

This paper investigated a new scheme for audio source separation: discrete, joint-state approaches. The theoretical contributions of this work include: the oracle estimator related to the proposed scheme; a significant 4dB SDR improvement in theoretical performance bounds compared to the generative GMM case; the raising question of the source correlation and consequences on the mixture states; a possible reduction of the computational cost of the decoding stage for a given performance level. An actual separation system is designed, showing a possible way of implementing the proposed scheme. However when the number of states increases, a small performance gain is reached due to an overfitting issue.

Future works may deal with the study of the source correlation in music and speech signals, and on the way of learning parameters appropriately in order to prevent overfitting. As the proposed method is valid for any number of sources, the evaluation may also be extended to more than two sources.

6. REFERENCES

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independant Component Analysis*. New York: J. Wiley, 2001.

[2] L. Benaroya and F. Bimbot, "Wiener based source separation with HMM/GMM using a single sensor," in *Proc. of ICA*, Nara, Japan, Apr. 2003, pp. 957–961.

[3] S. Roweis, "One Microphone Source Separation," *NIPS*, vol. 13, pp. 793–799, 2001.

[4] T. Virtanen, "Unsupervised Learning Methods for Source Separation in Monaural Music," in *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds. Springer-Verlag, 2006, pp. 267–296.

[5] R. Blouet, G. Rapaport, and C. Fevotte, "Evaluation of several strategies for single sensor speech/music separation," in *Proc. of ICASSP*, 2008, pp. 37–40.

[6] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proc. of ICA*. Paraty, Brazil: Springer, Mar. 2009.

[7] A. Ozerov, "Adaptation de modles statistiques pour la sparation de sources mono-capteur. Application la sparation voix / musique dans les chansons," Ph.D. dissertation, Univ. de Rennes 1, France, 2006.

[8] A. Ng and M. Jordan, "On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes." *NIPS*, vol. 14, 2001.

[9] E. Vincent, R. Gribonval, and M. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933–1950, 2007.

[10] T. Virtanen, "Monaural sound source separation by perceptually weighted non-negative matrix factorization," Tampere University of Technology, Tech. Rep., 2007.